Statistical inference course project

Phillip Rowe, 9/12/2019

Part I. Simulation Exercise

Overview

We will demonstrate two points of the Central Limit Theorem: 1) that the sample mean and sample variance estimate the theoretical mean and variance of the exponential probability distribution and 2)that the sample mean and sample variance have normal probability distributions.

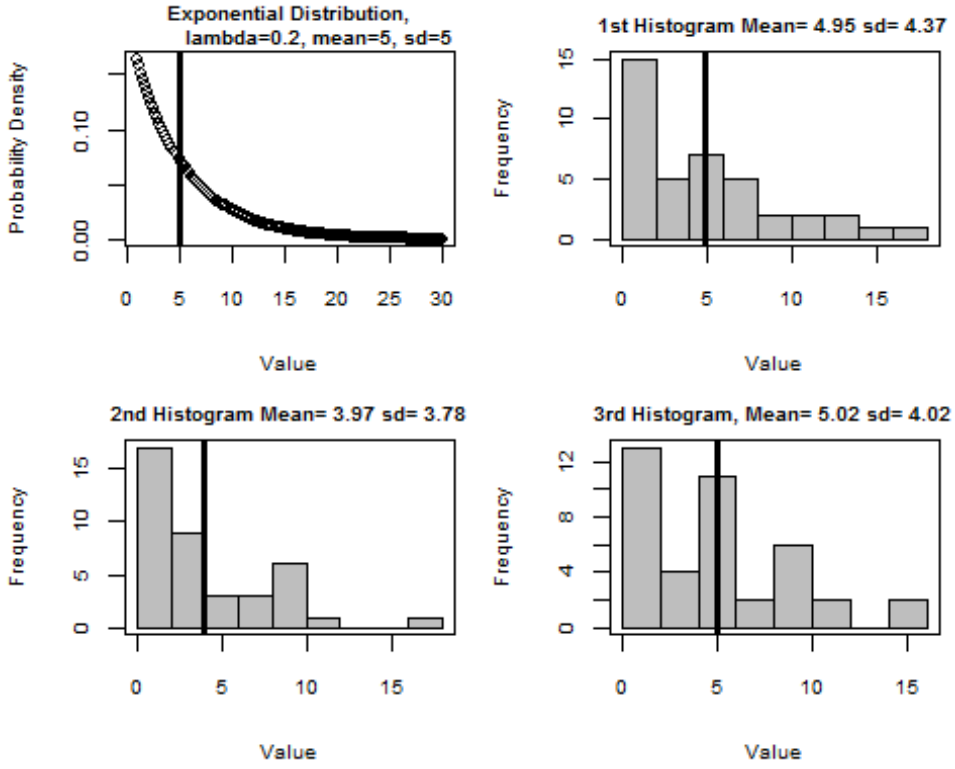
Simulations

The exponential distribution is given by f(x)=lambda*exp(-lambda*x), for x>=0, and its expected value (mean) and standard deviation are both equal to 1/lambda (see Figure 1 for plot of PDF). Samples can be generated in R using `rexp(n, lambda)`. We set lambda to 0.2 (mean = standard deviation = 5). In the code below, we create a dataframe of 1000 rows x 40 columns where each row is 40-samples of an exponential distribution. We then plot in Figure 1 three histograms of the first three rows of the dataframe.

```
library("dplyr");  library("plyr"); library(lubridate); library('reshape2')

lambda=0.2
# store 1000 simulations of 40 samples of distribution in a dataframe
alldata = NULL
for (i in 1 : 1000) alldata = rbind(alldata, rexp(40,lambda))
#calculate 1000 means, one for each of 40-sample rows
allmns=apply(alldata,1,mean); bigmean<-round(mean(allmns),2)
     #calculate 1000 standard deviations, one for each of 40-sample rows
allsds=apply(alldata,1,sd); bigsd<-round(mean(allsds),2)

par(mfrow=c(2,2),mai = c(.7, 0.7, 0.3, 0.1))
x<-seq(1,30,.2)
y<-lambda*exp(-(lambda*seq(1,30,.2)))
# a plot of the Exponential distribution with expected value 1/lambda
plot(x, y, ylab="Probability Density", xlab="Value",  main="Exponential Distribution,
          lambda=0.2, mean=5, sd=5",cex.main=.75, cex.lab=.75, cex.axis=.75)
abline(v=5,lwd=3)

mean1<-round(mean(alldata[1,]),2); sd1<-round(sd(alldata[1,]),2)
title<-paste("1st Histogram Mean=",mean1,"sd=",sd1)
hist(alldata[1,],10,ylab="Frequency",xlab="Value", main=title,cex.main=.75,
     col="gray",cex.lab=.75, cex.axis=.75)
abline(v=mean1,lwd=3); box()

mean2<-round(mean(alldata[2,]),2); sd2<-round(sd(alldata[2,]),2)
title<-paste("2nd Histogram Mean=",mean2,"sd=",sd2)
hist(alldata[2,],10,ylab="Frequency",xlab="Value", main=title,cex.main=.75,
     col="gray",cex.lab=.75, cex.axis=.75)
abline(v=mean2,lwd=3); box()

mean3<-round(mean(alldata[3,]),2)
sd3<-round(sd(alldata[3,]),2)
title<-paste("3rd Histogram, Mean=",mean3,"sd=",sd3)
```

```
hist(alldata[3,],10,ylab="Frequency",xlab="Value", main=title,cex.main=.75,
     col="gray",cex.lab=.75, cex.axis=.75)
abline(v=mean3,lwd=3); box()
```

Figure 1.

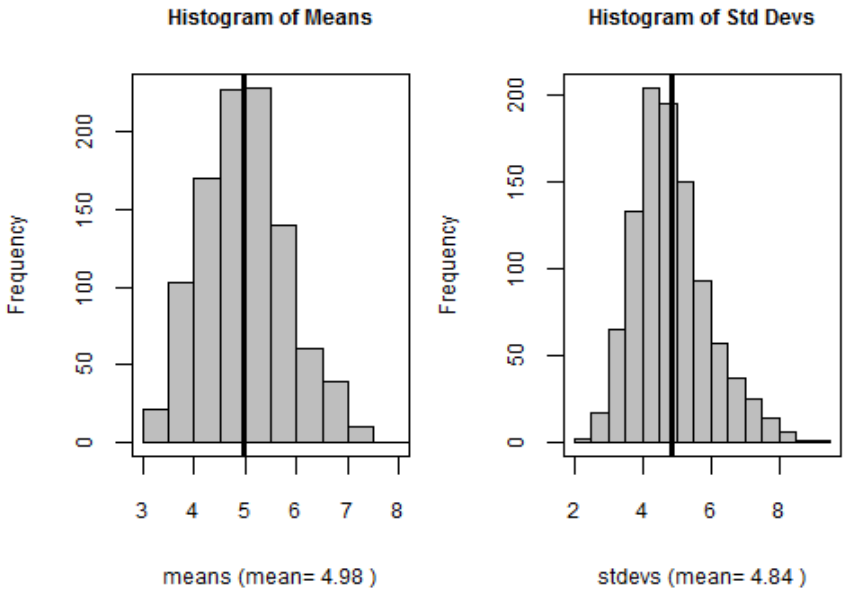


Sample Mean vs. Theoretical Mean; Sample Variance versus Theoretical Variance

Due to the random nature of the sampling and the relatively small sample size (n=40), some of these histograms approximate the PDF in the upper left corner better than others, and thus the sample mean and sample standard deviation shown on each histogram are only somewhat close to the theoretical values. That said, over the 1000 simulations, we calculated all the 40-sample means and standard deviations (stored in `allmeans` and `allsds`), and plotted their histograms, which appeared to be normal distributions in Figure 2. The mean of all 1,000 sample means was 4.98, and the mean of all 1,000 sample standard deviations was 4.84, both much closer to the theoretical values of the exponential distribution where lambda=0.2.

```
# now compare histograms of means and variances, which should look Gaussian
par(mfrow=c(1,2),mai = c(1, .8, .7, 0.1))
xlabel<-paste("means (mean=",bigmean,")")
hist(allmns, xlim=c(3,8), col="gray", ylab="Frequency",xlab=xlabel,
     main="Histogram of Means",cex.main=.75, cex.lab=.75, cex.axis=.75)
abline(v=bigmean,lwd=3); box()

hist(allsds,col="gray", xlab=paste("stdevs (mean=",bigsd,")"),ylab="Frequency",
     main="Histogram of Std Devs",
     cex.main=.75,
     cex.lab=.75, cex.axis=.75)
abline(v=bigsd,lwd=3); box()
```

Figure 2.



Part II. Basic Inferential Data Analysis

Assumptions

There were 60 different guinea pigs in the experimental data, so each dosage-supplement-length datapoint should be considered independent (i.e., unpaired). We use the t-test for our hypothesis tests, due to the relatively low number of samples per group (10). We assume that the sample mean length for tooth growth estimates the population mean of any guinea pig that would receive a simlar dosage. Variances are considered not equal for the t-tests.

Conclusions

1. **Ha: oj1 mean growth is more than 6 greater than oj_05 mean growth.** The 95% confidence interval is 6.21 to Infinity, so we reject the null hypothesis that the difference is less than 6. This difference represents a 6/13.23 = a 45% increase in growth due to the greater dosage.

2. **Ha: oj2 mean growth is greater than oj1 mean growth.** That said, the confidence interval is about +0.75, which is only about 3% increase over the mean growth of 22.7 for oj1 dosage.
3. **Ha: vc1 mean growth is greater than vc_05 mean growth.** The 95% confidence interval is 6.75 to Infinity, so we reject the null hypothesis that the difference is less than 6. This difference represents roughly a 6/7.98 = a 75% increase in growth due to the greater dosage.

4. **Ha: vc2 mean growth is greater than vc1 mean growth.** The difference is greater than 6, so we can again estimate that it is related to a 6/16.77~36% increase in growth.

5. **Ha: oj_05 mean growth is greater than vc_05 mean growth.** The difference at the 95% confidence level is only 1.7, however, and without a zero dosage control group for both supplements, it is not possible to make any conclusions for treatment based on this hypothesis test.

6. **Ha: oj1 mean growth is greater than vc1 mean growth.** The difference at the 95% confidence level is only 2.8, however, and the increase in growth vc1/vc_05 was greater at 75% vs. 45% for oj1/oj_05. Thus, it appears that vc is a more effective supplement to induce growth.

7. **H0: oj2 mean growth is the same as vc2 mean growth.** The confidence interval contains 0; thus, we fail to reject the null hypothesis.

Loading data and basic summary

We use boxplots in Figure 3 to summarize the experimental data. We observe that it appears the doubling of orange juice dosage (OJ) from 0.5 to 1.0 had a significant impact on tooth growth, but the increase of dosage from 1 to 2 seems to have a smaller impact. Contrastingly, the increase in dosage of ascorbic acid (VC) seems to have had a big impact on growth for both the 1 and 2 dosages. The mininum, maximum, mean, and standard error of each experiment is shown in the code block below. The dosages are 0.5, 1, and 2 mg/day.

Hypothesis testing

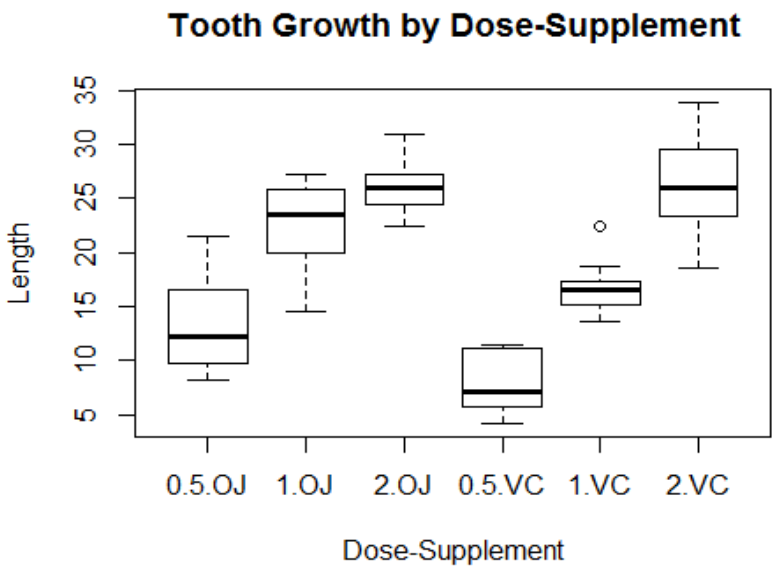
We used 4 t-tests to find the significance levels of differences in mean tooth length between successive dosages (see comments in code).

We also observe that the lower doses of OJ seem to spur more growth than VC, but at the highest dose, the growth appears similar. We performed three t-tests to compare means of the same dosage, different supplement.

Given the p-value is so small (4.4e-5), we reject the null hypothesis, opting for the alternative hypothesis that the oj2 sample has a mean that is larger than oj1. We could even pose a null hypothesis that the difference in means is 6 or less, and we would still reject the null hypothesis with an alpha of 5%, as the 95% confidence interval is 6.2 or greater.

```
library("dplyr"); library("plyr"); library('reshape2')
teeth<-group_by(ToothGrowth,dose,supp)

par(mfrow=c(1,1))
#png(file='teeth.png',width=480,height=480)
boxplot(len~dose+supp,teeth,main="Tooth Growth by Dose-Supplement",
   xlab="Dose-Supplement", ylab="Length")
```

Figure 3.



```r
dev.off()

teeth<-melt(data=ToothGrowth,id.vars=c("dose","supp"))
casted<-ddply(teeth, .(supp,dose), summarize, min=min(value), max=max(value),
              mean=mean(value), std_error=round(sd(value)/sqrt(10),2))
casted

##   supp dose  min  max  mean std_error
## 1   OJ  0.5  8.2 21.5 13.23      1.41
## 2   OJ  1.0 14.5 27.3 22.70      1.24
## 3   OJ  2.0 22.4 30.9 26.06      0.84
## 4   VC  0.5  4.2 11.5  7.98      0.87
## 5   VC  1.0 13.6 22.5 16.77      0.80
## 6   VC  2.0 18.5 33.9 26.14      1.52

oj_05<-subset(ToothGrowth,dose==0.5&supp=='OJ')$len
oj1<-subset(ToothGrowth,dose==1&supp=='OJ')$len
oj2<-subset(ToothGrowth,dose==2&supp=='OJ')$len
n<-10
vc_05<-subset(ToothGrowth,dose==0.5&supp=='VC')$len
vc1<-subset(ToothGrowth,dose==1&supp=='VC')$len
vc2<-subset(ToothGrowth,dose==2&supp=='VC')$len
#------------ ------------ ------------ ------------ ------------ ------------
# Hypothesis 0: mean tooth growth mu of dosage1 = mean growth of dosage 2
# Hypothesis A: len under dosage 2 is greater

#   ----------- 0.5 OJ vs 1 OJ  -----------
t.test(oj1,oj_05,alternative = 'greater',paired=FALSE)

##
##  Welch Two Sample t-test
##
## data:  oj1 and oj_05
## t = 5.0486, df = 17.698, p-value = 4.392e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.214316      Inf
## sample estimates:
## mean of x mean of y
##     22.70     13.23

#  ----------- 1.0 OJ  vs 2 OJ  -----------
t.test(oj2,oj1,alternative = 'greater',paired=FALSE)

## data:  oj2 and oj1
## t = 2.2478, df = 15.842, p-value = 0.0196
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7486236       Inf
## sample estimates:
## mean of x mean of y
##     26.06     22.70

#  ----------- 0.5 VC  vs 1 VC   -----------
t.test(vc1,vc_05,alternative = 'greater',paired=FALSE)

## data:  vc1 and vc_05
## t = 7.4634, df = 17.862, p-value = 3.406e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.746867      Inf
## sample estimates:
```

```
## mean of x mean of y
##     16.77      7.98

#  ----------- 1.0 VC vs 2 VC    -----------
t.test(vc2,vc1,alternative = 'greater',paired=FALSE)

## data:  vc2 and vc1
## t = 5.4698, df = 13.6, p-value = 4.578e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.346525      Inf
## sample estimates:
## mean of x mean of y
##     26.14     16.77

#------------ ------------ ------------ ------------ ------------ ------------
# Hypothesis 0: mean tooth growth of OC = mean growth of VC
# Hypothesis A: mean tooth growth is not equal under same dosage, different supplement
#   ----------- 0.5 OJ vs. 0.5 VC   -----------
t.test(oj_05,vc_05,alternative = 'two.sided',paired=FALSE)

## data:  oj_05 and vc_05
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean of x mean of y
##     13.23      7.98

#  ----------- 1 OJ vs. 1 VC  -----------
t.test(oj1,vc1,alternative = 'two.sided',paired=FALSE)

## data:  oj1 and vc1
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean of x mean of y
##     22.70     16.77

#   ----------- 2 OJ vs. 2 VC  -----------
t.test(oj2,vc2,alternative = 'two.sided',paired=FALSE)

## data:  oj2 and vc2
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```